Analytical validation of germline small variant detection using long-read HiFi genome sequencing

Nathan Hammond[1*], Linda Liao[1], Pun Wai Tong[1], Zena Ng[1], Thuy-Mi P. Nguyen[2], Chandler Ho[1], Yao Yang[1,2], Stuart A. Scott[1,2]

[1] Clinical Genomics Laboratory, Stanford Medicine, Palo Alto, CA 94304.
[2] Department of Pathology, Stanford University, Stanford, CA 94305.

Current affiliation:
[*] Influx Bio, San Francisco, CA 94124.

CORRESPONDENCE TO:

Stuart A. Scott, PhD
Professor,
Department of Pathology
Stanford University
Palo Alto, CA 94305
Telephone: 650-724-0973
Email: sascott@stanford.edu

SUPPLEMENTAL MATERIAL CONTENTS:

Supplemental Table S1. Short-read vs. long-read HiFi genome sequencing SNV/indel (<50 bp) accuracy in complex medically relevant genes (CMRG).

		Short-read	Long-read HiFi
SNVs	**Count**	17559	17559
	Recall	97.86%	98.42%
	Precision	99.33%	99.30%
Insertions (1-5bp)	**Count**	1518	1518
	Recall	92.29%	94.40%
	Precision	96.80%	93.79%
Insertions (6-15bp)	**Count**	204	204
	Recall	82.84%	87.75%
	Precision	90.31%	87.92%
Insertions (≥16bp)	**Count**	135	135
	Recall	74.81%	84.44%
	Precision	90.38%	90.98%
Deletions (1-5bp)	**Count**	1501	1501
	Recall	95.47%	97.40%
	Precision	96.12%	95.19%
Deletions (6-15bp)	**Count**	213	213
	Recall	91.55%	91.55%
	Precision	96.23%	91.82%
Deletions (≥16bp)	**Count**	146	146
	Recall	67.81%	87.67%
	Precision	91.43%	91.43%
All indels	**Count**	3605	3605
	Recall	95.86%	94.62%
	Precision	95.86%	93.76%
SNVs and indels	**Count**	21276	21276
	Recall	96.74%	97.73%
	Precision	98.71%	97.8.31%

indels: insertions/deletions; SNVs: single nucleotide variants.

Supplemental Table S2. Long-read HiFi genome sequencing SNV/indel (<50 bp) accuracy (mean ± standard deviation, n=3) versus average sequencing depth.

		5×	10×	15×	20×	25×
SNVs	**Count**	3298354 ± 58784	3298354 ± 58784	3298354 ± 58784	3298355 ± 58784	3298354 ± 58784
	Recall	78.1% ± 1.27%	97.0% ± 0.10%	99.3% ± 0.08%	99.6% ± 0.07%	99.7% ± 0.07%
	Precision	95.9% ± 0.28%	99.6% ± 0.00%	99.9% ± 0.01%	99.9% ± 0.01%	100.0% ± 0.01%
Insertions (1-5bp)	**Count**	209262 ± 25165	209262 ± 25165	209262 ± 25165	209262 ± 25165	209262 ± 25165
	Recall	65.5% ± 3.51%	90.9% ± 1.60%	96.6% ± 0.89%	98.3% ± 0.51%	98.9% ± 0.34%
	Precision	87.6% ± 1.53%	95.3% ± 0.76%	97.6% ± 0.50%	98.6% ± 0.30%	99.1% ± 0.23%
Insertions (6-15bp)	**Count**	22423 ± 2078	22423 ± 2078	22423 ± 2078	22423 ± 2078	22423 ± 2078
	Recall	61.8% ± 1.78%	89.7% ± 0.76%	96.2% ± 0.46%	97.9% ± 0.37%	98.7% ± 0.27%
	Precision	84.6% ± 0.99%	94.6% ± 0.52%	97.6% ± 0.24%	98.6% ± 0.27%	99.2% ± 0.17%
Insertions (≥16bp)	**Count**	4743 ± 924	4743 ± 924	4743 ± 924	4743 ± 924	4743 ± 924
	Recall	65.9% ± 0.93%	91.8% ± 0.39%	97.2% ± 0.20%	98.4% ± 0.19%	99.0% ± 0.36%
	Precision	89.4% ± 0.87%	96.6% ± 0.34%	98.4% ± 0.30%	99.0% ± 0.23%	99.3% ± 0.25%
Deletions (1-5bp)	**Count**	219611 ± 27525	219611 ± 27525	219611 ± 27525	219611 ± 27525	219611 ± 27525
	Recall	65.8% ± 3.15%	91.7% ± 1.28%	97.1% ± 0.74%	98.6% ± 0.39%	99.1% ± 0.27%
	Precision	87.7% ± 2.21%	95.4% ± 1.32%	97.8% ± 0.77%	98.7% ± 0.44%	99.1% ± 0.30%
Deletions (6-15bp)	**Count**	24404 ± 2987	24404 ± 2987	24404 ± 2987	24404 ± 2987	24404 ± 2987
	Recall	63.8% ± 2.86%	90.8% ± 0.80%	96.5% ± 0.60%	98.1% ± 0.30%	98.7% ± 0.24%
	Precision	85.1% ± 1.68%	94.9% ± 0.67%	97.6% ± 0.47%	98.5% ± 0.23%	98.9% ± 0.19%
Deletions (≥16bp)	**Count**	5272 ± 1276	5272 ± 1276	5272 ± 1276	5272 ± 1276	5272 ± 1276
	Recall	69.8% ± 2.92%	94.0% ± 0.79%	98.2% ± 0.37%	99.1% ± 0.16%	99.3% ± 0.12%
	Precision	91.4% ± 1.24%	97.6% ± 0.33%	98.9% ± 0.38%	99.3% ± 0.26%	99.5% ± 0.16%
All indels	**Count**	469971 ± 54387	469971 ± 54387	469971 ± 54387	469971 ± 54387	469971 ± 54387
	Recall	66.5% ± 3.08%	91.8% ± 1.23%	97.0% ± 0.68%	98.5% ± 0.37%	99.1% ± 0.26%
	Precision	87.4% ± 1.76%	95.3% ± 0.96%	97.7% ± 0.59%	98.7% ± 0.35%	99.1% ± 0.25%
SNVs and indels	**Count**	3784069 ± 109887	3784069 ± 109887	3784069 ± 109887	3784069 ± 109887	3784069 ± 109887
	Recall	76.4% ± 1.65%	96.2% ± 0.25%	98.9% ± 0.14%	99.5% ± 0.08%	99.6% ± 0.07%
	Precision	94.9% ± 0.53%	99.0% ± 0.16%	99.6% ± 0.10%	99.8% ± 0.06%	99.8% ± 0.05%

indels: insertions/deletions; SNVs: single nucleotide variants.

Supplemental Table S3. Long-read HiFi genome sequencing SNV/indel (<50 bp) concordance across library preparation methods (genome-wide).

		Genome Wide						RefSeq CDS	
		Low Complexity	Low Mappability	Segmental Duplications	All Difficult Regions	Not in Any Difficult Region	All	Not in Any Difficult Region	All
HM12878_BK3_2_Manual_Prep vs. HM12878	Count	1254505	505516	317613	1861931	3088444	4949218	14725	24351
	Concordance	90.06%	87.90%	90.59%	92.35%	99.87%	97.03%	99.69%	98.36%
UPN01_Swab_SRE vs. UPN01_Swab	Count	1238231	486664	310865	1840569	3088160	4927592	14722	24075
	Concordance	89.91%	87.96%	90.61%	92.25%	99.86%	97.00%	99.68%	98.29%
UPN02_Saliva_SRE_XS vs. UPN02_Saliva	Count	1233586	483127	303197	1826175	3077836	4902840	14904	24475
	Concordance	88.83%	87.18%	89.84%	91.47%	99.86%	96.72%	99.84%	98.12%
UPN02_Swab_SRE_XS_Miro_WithElutionBuffer vs. UPN02_Swab	Count	1248560	480405	307532	1849550	3124561	4972901	15062	24620
	Concordance	88.28%	87.34%	89.86%	91.07%	99.82%	96.56%	99.72%	98.31%
HM12878_Miro_C vs. HM12878	Count	1257853	488784	309435	1860799	3125290	4984873	15113	24239
	Concordance	89.61%	87.96%	90.71%	92.06%	99.86%	96.94%	99.79%	98.51%

UPN: Internal Unidentified Patient Number; RefSeq CDS: NCBI Reference Sequence gene coding sequence.

Supplemental Table S4. Long-read HiFi genome sequencing SNV/indel (<50 bp) accuracy across library preparation methods.

		Genome Wide						RefSeq CDS	
		Low Complexity	Low Mappability	Segmental Duplications	All Difficult Regions	Not in Any Difficult Region	All (High Confidence)	Not in Any Difficult Region	All (High Confidence)
HM12878 (used as reference)	**Count**	455730	184656	120104	889260	2848240	3737338	13998	20575
	Recall	99.07%	99.57%	99.44%	99.40%	99.92%	99.80%	99.85%	99.76%
	Precision	99.22%	99.83%	99.70%	99.54%	99.98%	99.87%	99.97%	99.96%
HM12878_BK3_2_ Manual_Prep	**Count**	455729	184656	120104	889259	2848240	3737337	13998	20575
	Recall	98.81%	99.44%	99.34%	99.23%	99.91%	99.75%	99.89%	99.78%
	Precision	98.99%	99.82%	99.74%	99.43%	99.97%	99.84%	99.98%	99.97%
HM12878_Miro_C	**Count**	455729	184656	120104	889259	2848240	3737337	13998	20575
	Recall	98.71%	99.36%	99.25%	99.16%	99.90%	99.72%	99.85%	99.78%
	Precision	98.92%	99.82%	99.77%	99.40%	99.98%	99.84%	99.99%	99.95%

RefSeq CDS: NCBI Reference Sequence gene coding sequence.

Supplemental Table S5. Long-read HiFi genome sequencing SNV/indel (<50 bp) reproducibility (genome-wide).

		Genome-Wide						RefSeq CDS	
		Low Complexity	Low Mappability	Segmental Duplications	All Difficult Regions	Not in Any Difficult Region	All	Not in Any Difficult Region	All
SNVs	**Count**	458126	402050	247696	1015567	2912745	3928324	14818	23709
	Concordance	88.60%	87.69%	90.72%	93.01%	99.86%	98.08%	99.66%	98.22%
Insertions (1-5bp)	**Count**	311980	26676	24215	330251	79513	410673	54	170
	Concordance	87.28%	76.00%	81.53%	87.76%	99.76%	90.12%	98.79%	94.01%
Insertions (6-15bp)	**Count**	51038	2906	3455	52755	8242	61088	6	77
	Concordance	82.69%	71.31%	75.49%	83.09%	99.59%	85.33%	100.00%	96.36%
Insertions (≥16bp)	**Count**	27720	4566	2287	28803	4761	33608	5	94
	Concordance	75.70%	65.44%	69.72%	76.10%	90.36%	78.12%	100.00%	82.88%
Deletions (1-5bp)	**Count**	308262	33403	23159	326574	79518	405348	64	198
	Concordance	90.65%	78.93%	83.74%	90.93%	99.81%	92.61%	99.47%	96.98%
Deletions (6-15bp)	**Count**	53522	3663	3717	55193	8291	63073	10	85
	Concordance	88.85%	77.08%	79.35%	89.05%	99.53%	90.35%	100.00%	97.49%
Deletions (≥16bp)	**Count**	23912	4235	1716	24404	3163	26506	6	68
	Concordance	88.47%	79.26%	78.16%	88.61%	97.73%	89.44%	96.97%	85.15%
All indels	**Count**	734544	73803	56650	776071	183428	958335	145	680
	Concordance	88.26%	76.89%	81.54%	88.64%	99.49%	90.65%	99.19%	93.13%
SNVs and indels	**Count**	1234559	477499	306244	1833547	3096232	4928621	14963	24401
	Concordance	88.26%	85.95%	88.92%	90.97%	99.84%	96.52%	99.66%	98.07%

indels: insertions/deletions; RefSeq CDS: NCBI Reference Sequence gene coding sequence; SNVs: single nucleotide variants.

Supplemental Table S6. Long-read HiFi genome sequencing SNV/indel (<50 bp) paired specimen concordance (genome-wide).

		Genome-Wide						RefSeq CDS	
		Low Complexity	Low Mappability	Segmental Duplications	All Difficult Regions	Not in Any Difficult Region	All	Not in Any Difficult Region	All
Blood vs Saliva	**Count**	1239818	480919	306199	1837182	3101144	4937145	24429	15086
	Concordance	89.13%	87.49%	90.35%	91.70%	99.85%	96.79%	98.25%	99.72%
Blood vs Swab	**Count**	1245625	486775	308450	1844823	3101488	4945105	24524	15099
	Concordance	89.87%	88.01%	90.90%	92.25%	99.87%	97.01%	98.32%	99.78%
Swab vs Saliva	**Count**	1476535	707260	356983	2109841	3106447	5215082	25522	15099
	Concordance	86.13%	82.60%	89.15%	89.22%	99.87%	95.49%	97.49%	99.73%

RefSeq CDS: NCBI Reference Sequence gene coding sequence.

Supplemental Table S7. Sample summary and quality metrics for full-depth HiFi genome sequencing samples.

Sample Name	Depth (×)	Total Reads	Mean Read Length	Q30 Bases (Yield)	Aligned Reads	Mismatch Rate	indel Rate	Total SNVs	Total indels	Ti/Tv
HM12878 [a]	32.78	6119036	16834	99654688439	6081335	0.21%	0.22%	3940265	856248	1.97
HM24385 [a]	27.08	5250688	16290	82525012315	5197613	0.24%	0.23%	3933551	852669	1.98
HM24143 [a]	33.59	6053415	17460	1.01653E+11	6010254	0.22%	0.24%	3996962	863716	1.97
HM24149 [a]	30.95	6415079	15179	94063731893	6371689	0.23%	0.24%	3911716	843108	1.98
HM24631 [a]	29.22	5424725	17011	88928721744	5368886	0.23%	0.24%	3907195	851021	1.97
HM24694 [a]	30.75	5699699	17051	93352599087	5639241	0.23%	0.26%	3891630	839764	1.97
HM24695 [a]	33.33	6353081	16529	1.00995E+11	6307031	0.24%	0.26%	3945732	850240	1.96
UPN01_Blood [a]	35.09	6787711	16345	1.07053E+11	6717756	0.26%	0.24%	3928508	855811	1.96
UPN01_Saliva [a]	28.41	6910650	14927	1.00171E+11	5968237	0.26%	0.21%	3915612	850776	1.97
UPN01_Swab [a]	29.08	6186531	16104	96103465104	5663852	0.26%	0.25%	3914032	848269	1.97
UPN02_Blood [a]	30.78	7994583	12126	94323437456	7929472	0.24%	0.21%	3976426	859362	1.97
UPN02_Saliva [a]	21.63	8039806	10145	79339979547	6943528	0.24%	0.21%	3946229	846260	1.98
UPN02_Swab [a]	25.74	8235127	10498	84372724958	7728556	0.24%	0.19%	3962033	855282	1.98
UPN02_Saliva_SRE_XS [a]	24.71	8321585	11313	91818488308	7047888	0.23%	0.18%	3959309	855444	1.98
UPN02_Swab_SRE_XS_ Miro_WithElutionBuffer [a]	28.32	8105789	11878	93605919185	7515551	0.24%	0.21%	3968366	858376	1.98
UPN01_Swab_SRE [a]	28.22	7546392	12790	93660217330	6984738	0.24%	0.22%	3903339	846912	1.97
HM12878_BK3_2_ Manual_Prep [a]	31.14	5151792	19074	94490166479	5108108	0.24%	0.25%	3943356	851696	1.97
HM12878_Miro_C [a]	27.87	6170618	14248	85173043165	6114838	0.23%	0.22%	3924021	848806	1.97
NA12878_rep2 [b]	28.79	9080816	9963	87340101054	9022556	0.22%	0.21%	3906222	851259	1.98
NA12878_rep3 [b]	24.60	4442992	17435	74191449578	4414246	0.23%	0.29%	3907591	822481	1.98
NA24385_rep2 [b]	52.59	11678114	14190	1.61069E+11	11572448	0.22%	0.21%	3961972	861871	1.97
NA24385_rep3 [b]	31.61	8851868	11237	96200017899	8777876	0.22%	0.20%	3931942	852587	1.98
Average	**30.29**	**7037277**	**14483**	**95458372006**	**6749350**	**0.24%**	**0.23%**	**3935273**	**850998**	**1.9732**
Stdev	**5.94**	**1654707**	**2781**	**16610469493**	**1586860**	**0.012%**	**0.027%**	**27449**	**8599**	**0.0065**
Suggested QC Thresholds	**>25**	**>2900000**	**>7500**	**>76000000000**	**>3600000**	**0.20-0.27%**	**0.15-0.31%**	**>3850000**	**>825000**	**1.95-1.99**

UPN: Internal Unidentified Patient Number; SNVs: single nucleotide variants.
[a] HiFi genome sequencing data available at the NCBI BioProject database (see **Data Access**); [b] Publicly available HiFi genome sequencing data (see **Materials and Methods**).

Supplemental Table S8. Sample summary and quality metrics for low-depth and downsampled HiFi genome sequencing samples.

Sample Name	Depth (×)	Total Reads	Mean Read Length	Q30 Bases (Yield)	Aligned Reads	Mismatch Rate	indel Rate	Total SNVs	Total indels	Ti/Tv
HM24385_5X	4.87	943325	16293	14830179142	933798	0.24%	0.23%	3104302	609202	2.02
HM24385_9X	8.92	1730415	16288	27194219152	1712968	0.24%	0.23%	3721212	764312	2.01
HM24385_10X	10.01	1940826	16288	30501566323	1921362	0.24%	0.23%	3778375	781184	2.00
HM24385_15X	14.88	2886209	16289	45359635045	2857283	0.24%	0.23%	3878285	820167	1.99
HM24385_20X	20.03	3884562	16289	61051709589	3845421	0.24%	0.23%	3909755	839123	1.98
HM24385_25X	24.91	4830016	16289	75910092968	4781358	0.24%	0.23%	3926747	849427	1.98
HM12878_5X	4.93	919568	16836	14977178904	913930	0.21%	0.22%	3112290	611260	2.02
HM12878_9X	8.87	1654201	16839	26946351087	1644083	0.21%	0.22%	3702094	759353	2.01
HM12878_10X	10.17	1897856	16840	30918676150	1886268	0.21%	0.22%	3771680	779146	2.00
HM12878_15X	15.08	2814260	16838	45841585651	2797059	0.21%	0.22%	3870735	817488	1.99
HM12878_20X	20.0	3732596	16837	60798430047	3709748	0.21%	0.22%	3902077	835375	1.98
HM12878_25X	24.92	4651397	16837	75761835414	4622879	0.21%	0.22%	3920444	846505	1.98
HM24631_5X	4.98	923679	17014	15144083222	914112	0.23%	0.24%	3149290	618287	2.01
HM24631_9X	9.06	1682128	17014	27579998400	1664949	0.23%	0.24%	3703364	762827	2.00
HM24631_10X	9.94	1844971	17012	30246883500	1826055	0.23%	0.24%	3745601	776073	2.00
HM24631_15X	14.9	2767351	17012	45370872861	2738658	0.23%	0.24%	3844940	815622	1.99
HM24631_20X	19.87	3689388	17011	60482927715	3651275	0.23%	0.24%	3875221	834253	1.98
HM24631_25X	25.12	4665130	17010	76475932583	4617078	0.23%	0.24%	3895568	845902	1.97
HM12878_Miro_mi22030208 [a]	9.26	2286438	12774	28314538817	2267016	0.24%	0.22%	3730921	770714	2.00
HM12878_Miro_mi21070127 [a]	8.96	2171460	13015	27394118511	2152883	0.24%	0.22%	3707170	761419	2.00

SNVs: single nucleotide variants.

[a] HiFi genome sequencing data available at the NCBI BioProject database (see **Data Access**).

Supplemental Table S9. Sample summary and quality metrics for full-depth short-read genome sequencing samples.

Sample Name	Depth (×)	Total Reads	Read length	Q30 Bases (Yield)	Aligned Reads	%Mismatched bases R1	%Mismatched bases R2	Total SNVs	Total variants	Ti/Tv
NA12878_SR_40X	40.27	867741180	150	115678371270	798780802	0.32	0.58	3956523	4888777	1.98
NA24385_SR_40X	40.39	512689412	250	114519736312	471570488	0.53	0.83	3977090	4882675	1.96
NA24149_SR_40X	40.48	511578792	250	112688350774	474542534	0.53	0.92	3954675	4854345	1.97
NA24143_SR_40X	40.33	521163374	250	116117493757	481629809	0.51	0.84	4018723	4930621	1.96
NA24631_SR_40X	40.37	525325580	250	114865739238	473137009	0.55	0.97	3962506	4874074	1.97
NA24694_SR_40X	40.06	856909424	150	111953568851	777964484	0.40	0.74	3912575	4839894	1.99
NA24695_SR_40X	40.21	877375854	150	111692131363	797980661	0.37	0.83	3953589	4888047	1.99
Average	**40.30**	**667540517**		**113930770224**	**610800827**	**0.46**	**0.82**	**3962240**	**4879776**	**1.97**
Stdev	**0.14**	**187049921**		**1803762394**	**169265244**	**0.09**	**0.13**	**31734**	**28905**	**0.013**

SNVs: single nucleotide variants.

Supplemental Table S10. Sample summary and quality metrics for low-depth and downsampled short-read genome sequencing samples.

Sample Name	Depth (×)	Total Reads	Read length	Q30 Bases (Yield)	Aligned Reads	%Mismatched bases R1	%Mismatched bases R2	Total SNVs	Total variants	Ti/Tv
NA24385_SR_5X	5.16	64539114	250	14416401311	60313462	0.53	0.83	2953605	3476643	2.01
NA24385_SR_10X	10.58	132599146	250	29619147442	123610141	0.53	0.83	3698944	4436940	1.99
NA24385_SR_15X	15.49	194465316	250	43438586325	180885572	0.53	0.83	3877461	4687762	1.98
NA24385_SR_20X	20.37	256354030	250	57263108330	237932567	0.53	0.83	3939966	4796567	1.98
NA24385_SR_25X	25.24	318219590	250	71081605599	294704521	0.53	0.83	3958021	4834558	1.97
NA24385_SR_30X	30.77	388934272	250	86877151094	359292524	0.53	0.83	3968254	4858794	1.97
NA24385_SR_35X	35.59	450816766	250	100699251108	415556503	0.53	0.83	3973559	4872317	1.97

SNVs: single nucleotide variants.